Mathematical Biosciences and Engineering

http://www.aimspress.com/journal/mbe

MBE, 20(6): 10444–10458.
DOI: 10.3934/mbe.2023459
Received: 29 December 2022
Revised: 09 March 2023
Accepted: 20 March 2023
Published: 06 April 2023

Research article

Prediction Model of hospitalization time of COVID-19 patients based on Gradient Boosted Regression Trees

Zhihao Zhang[1,†], Ting Zeng[1,2,†], Yijia Wang[3], Yinxia Su[1], Xianghua Tian[1], Guoxiang Ma[1], Zemin Luan[2] and Fengjun Li[1,*]

[1] College of Medical Engineering and Technology, Xinjiang Medical University, Urumqi 830017, China

[2] School of Public Health, Xinjiang Medical University, Urumqi 830017, China

[3] College of Mathematics and System Science, Xinjiang University, Urumqi 830017, China

[†] These authors contributed equally.

* **Correspondence:** Email: zwhlfj@xjmu.edu.cn.

Abstract: When an outbreak of COVID-19 occurs, it will cause a shortage of medical resources and the surge of demand for hospital beds. Predicting the length of stay (LOS) of COVID-19 patients is helpful to the overall coordination of hospital management and improves the utilization rate of medical resources. The purpose of this paper is to predict LOS for patients with COVID-19, so as to provide hospital management with auxiliary decision-making of medical resource scheduling. We collected the data of 166 COVID-19 patients in a hospital in Xinjiang from July 19, 2020, to August 26, 2020, and carried out a retrospective study. The results showed that the median LOS was 17.0 days, and the average of LOS was 18.06 days. Demographic data and clinical indicators were included as predictive variables to construct a model for predicting the LOS using gradient boosted regression trees (GBRT). The MSE, MAE and MAPE of the model are 23.84, 4.12 and 0.76 respectively. The importance of all the variables involved in the prediction of the model was analyzed, and the clinical indexes creatine kinase-MB (CK-MB), C-reactive protein (CRP), creatine kinase (CK), white blood cell count (WBC) and the age of patients had a higher contribution to the LOS. We found our GBRT model can accurately predict the LOS of COVID-19 patients, which will provide good assistant decision-making for medical management.

Keywords: machine learing; COVID-19; length of stay; predictive analytics; GBRT

1. Background

A novel coronavirus COVID-19 pandemic was spread worldwide in late 2019. COVID-19 belongs to the mRNA virus, which is initially characterized by fever, fatigue, cough, muscle pain, etc. some patients will have severe symptoms such as acute respiratory distress, septic shock, metabolic acidosis, and even organ failure, resulting in increased mortality [1]. As of March 28, 2022, the cumulative confirmed cases have reached 301,988,196, the cumulative deaths have reached 5,477,778, and the case fatality rate is 2.08%. COVID-19 virus has high concealment and can spread rapidly among people. when an outbreak of COVID-19 occurs, it will quickly bring a great burden on local medical resources, resulting in a significant increase in the demand for hospital beds and a shortage of medical equipment [2]. Yaesoubi [3] developed tools for early warning of COVID-19 hospitalization overload. The tool provided simple and easy-to-communicate decision rules to predict whether local hospital occupancy is expected to exceed capacity within a 4- or 8-week period if no additional mitigating measures are implemented. Chen et al. [4] took into account disease progression and changes in biomarkers over time and modeled them using historical regression trees (HTREEs). The study identified important biomarkers associated with the prognosis of COVID-19 patients, characterized the time-to-event process and obtained dynamic predictions at the individual level.

At present, scientists and doctors are actively exploring the diagnosis and treatment methods of this disease. A lot of research work has also been done in the field of computer-aided diagnosis, including the model for predicting the risk of COVID-19 in the general population, the model for diagnosing suspected infected patients with COVID-19, and the prognosis model for patients with COVID-19 [5]. Bardelli [6] analyzed different parameters related to the personal evolution of COVID-19 (i.e., time of recovery, length of stay in hospital and delay in hospitalization). A Bayesian Survival Analysis was performed considering the age factor and period of the epidemic as fixed predictors to understand how these features influence the evolution of the epidemic. You et al. [7] evaluated the significance of diaphragm thickness (DT) in assessing the nutritional status and predicting the length of hospital stay (LOS) of patients with COVID-19. According to the model of multiple linear regression analysis, the DT at admission and mechanical ventilation were independent risk factors that contributed to LOS. Some scholars have analyzed the importance of the clinical indicators that affect LOS in hospitals. Chiam et al. [8] based on the clinical data of 58 patients with COVID-19 in the second affiliated Children's Hospital and Yuying Children's Hospital of Wenzhou Medical University, through epidemiological statistical analysis, it was found that patients with overweight / obesity and abnormal liver function were more likely to prolong LOS. Usher et al. [9] took the patient data of 36 hospitals in the Midwest and north of the United States as the research objects, and constructed an analysis model for sharing unidentified patient data across systems. The prediction results showed that the median LOS was 5.0 days, with an average of 8.2 days. Consistent predictors of LOS included age, critical illness, oxygen demand, weight loss, and nursing home admission. The chest CT scoring system Orebro COVID-19 Scale (OCoS) is implemented in the clinical routine in Orebro, Sweden. The system scores according to the degree of lung involvement. Ahlstrand et al. [10] evaluated the correlation between the CT score at admission and intensive care and the LOS in hospital and intensive care, and compared it with C-reactive protein and lymphocyte count, The results showed that the predictive effect of OCoS score was better than that of basic inflammatory biomarkers. Lasbleiz [11] study aim was to compare phenotypic characteristics between in and outpatients with diabetes and infected by COVID-19 and to build an easy-to-use hospitalization prediction risk score. DIAB score is an easy-to-use score integrating five variables to help clinicians better manage patients with DM and avert the saturation of emergency care units.

Some other scholars have predicted the LOS of COVID-19 patients according to demographic and clinical indicators, and some studies assessed the reasons for patients' prolonged LOS. The data of 1099 COVID-19 patients from Chinese mainland hospitals in 30 provinces, autonomous regions and municipalities directly under the Central Government showed that the median hospital stay of all patients was 12.0 days (average 12.8 days) [12]. A discrete-time model was developed by [13] Leclerc et al. to examine the impact of using bed paths or predicting bed occupancy rates only based on the average LOS of bed types. After comparing the bed occupancy rates predicted by COVID-19's model in England and the publicly available bed occupancy data between March and August 2020, it is found that LOS has regional heterogeneity and the national average LOS of COVID-19 may not be suitable for local. Ebinger et al. [14] developed three machine learning algorithms to predict the possibility of long-term LOS (defined as 8 days), to provide a basis for hospital bed demand decision-making, and to help clinicians answer COVID-19 patients' consultation about LOS. Li et al. [2] taking 97 patients from Beijing You'an Hospital from January 21, 2020, to March 21, 2020, as the research object, using the multivariable Cox proportional hazards regression method based on the minimum Chichi information standard value, a nomogram was constructed for demographic and clinical variables. The results showed that the model can accurately predict the LOS of COVID-19 patients. Mahboub et al. (2021) [15] took the clinical data of 2017 COVID-19 cases reported by the Dubai health authority as the research object, and established a decision tree (DT) model to predict the LOS of COVID-19. The model showed good performance, the determination coefficient R2 was 49.8%, and the median absolute deviation was 2.85 days. Lopez cheda, Jacob, Cao, and de Salazar [16] established a non-parametric mixed treatment model based on the COVID-19 epidemic detection data in Galicia, Spain, to evaluate the LOS in the hospital ward (HW) and Intensive Care Unit (ICU); Monte Carlo algorithm was used to simulate the demand of COVID-19 hospitals. They found gender and age were the key to accurate prediction of the model. Henzi et al. [17] based on the data of 557 critically ill patients with COVID-19 in Switzerland, according to the variables within 24 hours after admission to the intensive care unit, developed a semi-parametric distribution index model to predict the individual LOS of patients. Dan et al. [18] studied 733 patients in Wuhan, China before March 18, 2020. Based on demographic, clinical and laboratory data, a prediction model of ICU LOS of survivors based on the least absolute shrinkage and selection operator (LASSO) penalty was established. Qi et al. [19] used machine learning method based on CT radiological data to predict the LOS of patients with pneumonia associated with SARS-COV-2 infection. Rozenbaum [20] developed a decision tool that can provide explainable and patient-specific prediction of in-hospital mortality and LOS for COVID-19-positive patients. The model can aid healthcare systems in bed allocation and distribution of vital resources.

Ensemble learning algorithms have made great achievements in various research fields, and we noticed that the GBRT model has a large number of applications in the biomedical field. In this paper, it is suggested that in the management of infectious disease hospitals, a prediction of the discharge time of patients should be added to show the expected time of each patient. By doing so, patients can be encouraged psychologically and managers can make auxiliary decisions. We aimed to use the GBRT algorithm to establish a prediction model for predicting the LOS of COVID-19 patients based on the demographic and clinical index data of 166 patients, so as to provide a basis for relevant health departments to accurately predict the LOS of COVID-19 patients. In addition, we analyzed the importance of related variables to the model prediction.

2. Materials and methods

2.1. Data sources

The subjects of this study were the COVID-19 patients of a hospital in Urumqi, Xinjiang from July 19, 2020 to August 26, 2021. Their data were collected through the admission medical record management system, including the demographic characteristics and clinical indicators. Variables include gender, age, current medical history, past history, epidemiological contact history, smoking history, drinking history and family history, and the relevant examination results of the patient's first visit in the fever clinic or the first (or early) after admission, such as RT-PCR detection of viral nucleic acid, blood routine, urine routine, stool routine, liver and kidney function of pharyngeal swab and sputum sars-cov-2 Electrolyte, CRP, interleukin-6 (IL-6), procalcitonin (PCT), erythrocyte sedimentation rate (ESR), blood glucose, coagulation, lactate dehydrogenase (LDH), myocardial zymogram, myoglobin, troponin (TNI), electrocardiogram whole chest film or lung imaging, etc.

There were 166 patients with COVID-19 in the data set, including 75 males and 91 females. The symptom types of patients included 30 asymptomatic patients, 33 mild patients, 103 ordinary patients and no severe and critical patients. Discrete features of symptoms or medical history were counted, excluding the count result as 0 features. Table 1 shows the symptoms or medical history of some patients. As the clinical examination data of different patients are generated at different time points, the data recording is irregular in the time dimension. In addition, due to the different data categories of each clinical examination, the data recorded in the characteristic dimension is irregular. In order to ensure the reliability of data, we adopted the method of deleting missing records. Records with missing clinical indicator values were excluded clinical indicators. Data of patients younger than 18 years were excluded.

2.2. Methods

2.2.1. GBRT

GBRT is a boosting type of ensemble learning algorithm. Ensemble learning is a technical framework. It combines multiple different base models to complete the corresponding work in order to achieve more efficiency and accuracy. At present, the commonly used integrated learning frameworks include bagging, boosting and stacking. The boosting framework uses multiple groups of base models for training respectively, and the results of all base models are linearly combined to obtain more robust prediction results. Figure 1 is a schematic diagram of boosting the ensemble learning framework.

The overall model based on the boosting framework can be described by a linear combination:

$$F(x) = \sum_{i}^{m} h_i(x) \tag{1}$$

Where $h_i(x)$ represents the base model. The training goal of the overall model is to make the predicted value $F(x)$ approach the real value y. experts and scholars use the idea of greedy algorithm

Table 1. Partial statistical examples of symptoms and medical history of patients (Gender: 1 = male, 2 = female, disease outcomes: 1 = negative, 2 = positive nucleic acid, types: 1 = asymptomatic 2 = mild, 3 = ordinary).

No	Gender	Age	Fever	Cough	Sputum	Chest tightness	Myalgia	Sore throat	Dry mouth	Headache	Fatigue	Diarrhea	Nausea and vomingting	Hypertension	Diabetes	Coronary heart disease	Chronic bronchitis	Disease outcomes	Types	LOS
1	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	16
2	1	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	12
3	1	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	19
4	2	9	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	18
5	1	11	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	12
90	2	34	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	14
91	1	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	27
92	1	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	17
93	1	35	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	17
94	1	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	2	19
95	2	35	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	12
96	1	36	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	3	13
160	2	56	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	3	31
161	2	57	0	1	0	1	0	0	0	0	1	1	0	1	0	0	0	1	3	13
162	1	57	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	3	14
163	2	59	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	3	18
164	1	66	0	1	1	0	0	0	0	0	0	0	0	1	0	0	1	1	3	19
165	2	74	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	3	25
166	2	76	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	1	3	19

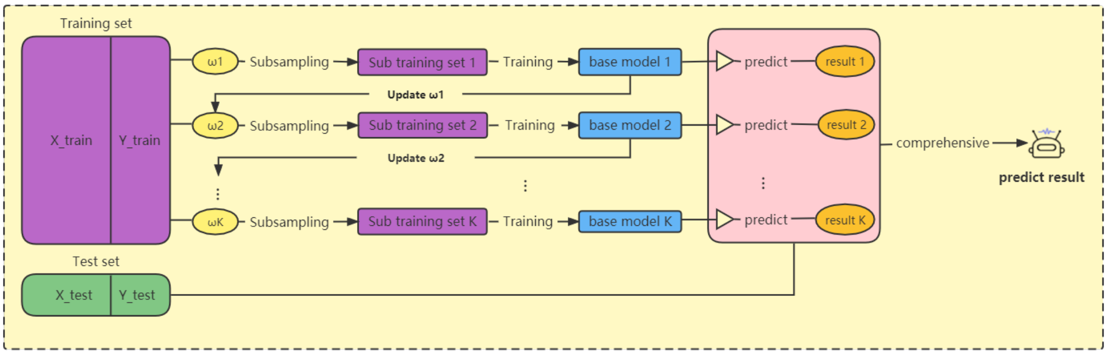

Figure 1. Schematic diagram of boosting integrated learning framework.

to make each base model undertake departmental prediction tasks and approach their own prediction tasks respectively, and focus on overcoming the errors generated by each base model.

$$F^{i}(x) = F^{i-1}(x) + h_{i}(x) \tag{2}$$

Fit the residual. Introduce an arbitrary loss function and fit the inverse gradient.

$$F^{i}(x) = F^{i-1}(x) + \arg\min \sum_{j}^{n} L(y_{j}, F^{i-1}(x_{j})) + h_{i}(x_{j}) \tag{3}$$

GBRT is a boosting integrated learning model based on tree structure. For the M features of a given n records, K tree functions are used to predict the output:

$$\hat{y}_{i} = \phi(x_{i}) = \sum_{K=1}^{K} f_{k}(x_{i}), f_{k} \in \Gamma \tag{4}$$

$$\Gamma = \{f(x) = w_{q(x)}\}(q : \Re^{m} \to T, \omega \in \Re^{T}) \tag{5}$$

Where q represents the structure of each tree mapping records to corresponding leaf indexes; T is the number of leaves on the tree; Each f corresponds to an independent tree structure q and leaf weight w; w_i represents the score on the ith leaf. The value of the leaf node region is estimated by linear search to minimize the loss function, and then the regression tree is updated.

2.2.2. GBRT-based prediction model of the length of stay for COVID-19

Data preprocessing

Generally, the examination data of 166 patients can be used to construct 166 data records based on the examination data on the day of admission for the prediction of length of stay. However, this approach does not make good use of clinical examination data generated during a patient's hospitalization. In order to maximize the utilization of data records, we used the data of clinical indicators examination results of patients during hospitalization to expand the data sample.

The data included 19 inherent characteristics of the patients, including gender, age, admission symptoms and medical history, as well as 9 clinical indicators obtained through examination after admission.

The clinical data varied at different time points, but the 19 inherent characteristics remained constant. Therefore, the data set can be populated by copying inherent features. The predicted target feature of the data set is the LOS, and the discharge time is known, so the LOS corresponding to the data records at different time points can be calculated.

A total of 3,141 records were generated from 166 patients in the original data. After the deletion of missing records, 11 patients had no valid records and 14 patients were younger than 18 years of age. The final data set was constructed from 324 records of 141 patients.

Splitting of the data set

In order to avoid information leakage in the training process, 141 patients were divided into the training set and test set data, and random sampling was conducted according to the ratio of 8 : 2. Then, through statistical checks, the number of data set records is also controlled to about 8 : 2.

Data regularization

In order to make the data more regular and convenient for model training, we have carried out data regularization. The formula is as follows:

$$L(\phi) = \sum_i l(y_i \hat{;} y_i) + \sum_k \Omega(f_k) \tag{6}$$

$$\Omega(f) = \gamma T + \frac{1}{2}\lambda\|\omega\|^2 \tag{7}$$

Where L is a differentiable convex loss function to measure the difference between prediction l and target y_i ; The second term $\Omega(f)$ penalizes the complexity of the model. Smoothing learning weights can avoid over fitting.

Data standardization

The distribution range of data features varies greatly. In order to accelerate the convergence speed of the model, min-max standardization is adopted to scale the features to between 0 and 1. The formula is as follows:

$$X = \frac{x - \min}{\max - \min} \tag{8}$$

Where max represents the maximum value of the feature in the sample data and min represents the minimum value of the feature in the sample data. x represents raw data, and X represents normalized data.

3. Results

3.1. Basic information about the research objects

In order to intuitively describe the data, we made statistical tables for age and length of hospital stay, 9 clinical indicators, 15 discrete features of symptoms or disease history. According to the data in Table 2, we can find that the age range of patients is 18–76 years old, the length of hospital stay is 8–33 days, the median length of hospital stay is 17 days, and the average of LOS is 18.06 days. Compared with the literature [12], the median LOS was 12.0 days and the average LOS was 12.8 days, the LOS collected by us was relatively large.

Table 2. Characteristics of age and LOS.

	Age (n = 141)	LOS (n = 141)
mean	36.73	18.06
min	18.00	8.00
25%	26.00	14.00
50%	35.00	17.00
75%	46.75	20.00
max	76.00	33.00

There are other features of our data table entries. Some entries are completely empty, so we do not make statistical description. The data in Table 3 showed that most patients had inflammatory pathologic features including cough, fever, sore throat and fatigue.

Table 3. Statistics of patients corresponding to discrete characteristics.

Category characteristics	Symptoms or disease history count (n = 141)
Fever	17 (12.0%)
Cough	21 (14.8%)
Sputum	5 (3.5%)
Chest tightness	3 (2.1%)
Myalgia	3 (2.1%)
Sore throat	15 (10.6%)
Dry mouth	1 (0.7%)
Headache	2 (1.4%)
Fatigue	23 (16.3%)
Diarrhea	5 (3.5%)
Nausea and vomingting	1 (0.7%)
Hypertension	15 (10.6%)
Diabetes	7 (4.9%)
Coronary heart disease	2 (1.4%)
Chronic bronchitis	4 (2.8%)

Table 4 shows the data records of 324 clinical examinations of 141 patients during hospitalization. We have counted the mean, maximum, minimum and three quantiles of each indicator.

3.2. Construction and evaluation of LOS model of COVID-19 patient

The Hyperparameters of the GBRT include: the learning rate, the number of estimators, the maximum depth of the tree, the number of split nodes in the sample, the minimum sample required for the leaf nodes, and the loss function. GridResearchCV was used to automatically find the optimal hyperparameters, and 10-fold cross-validation was sampled for training, where the loss function is fixed as

Table 4. Statistics of clinical index results.

	WBC	LY	CRP	ALT	AST	Cr	LDH	CK	CK-MB
mean	6.52	2.30	3.02	25.66	18.24	78.43	198.83	69.46	12.71
min	2.00	0.49	0.01	5.00	5.00	28.00	75.40	7.06	1.35
25%	5.12	1.48	0.43	13.00	12.00	65.00	165.48	44.54	7.97
50%	6.20	1.93	1.22	10.00	16.00	76.00	190.79	58.36	10.22
75%	7.83	2.37	3.59	29.00	21.00	90.00	221.63	78.43	13.54
max	14.99	36.10	88.59	181.00	92.00	181.00	349.97	379.42	152.06

WBC(10^9/L) represents the number of white blood cells, LY(10^9/L) represents the number of lymphocytes, CRP(mg/L) represents the number of reactive proteins, ALT(U/L) represents alanine aminotransferase, AST(U/L) represents aspartate aminotransferase, Cr(umol/l) represents creatinine, LDH(U/L) represents lactate deaminase, CK(U/L) represents creatine kinase, and CK-MB(U/L) represents creatine kinase isozyme.

squared error. In order to find the optimal super parameters more stably, we looked for five groups of super parameter candidates. The parameter settings are shown in Table 5.

Table 5. Hyperparameters value of the GBRT model.

Hyperparameters	Model1	Model2	Model3	Model4	Model5
learning rates	0.01	0.01	0.01	0.01	0.01
the depth of the tree	2	2	2	2	2
min samples split	6	2	8	7	6
min samples leaf	3	8	2	2	2
number of estimators	160	70	70	70	80

According to the hyperparameters results in Table 5, the GBRT model was trained. The data set had 359 records and 28 features, which were divided into training set and test set according to 8 : 2. After the model training, input 2the data characteristics of the test set to predict the discharge time. The prediction results of the model on the test set are shown in Table 6.

Table 6. Prediction results of GBRT model.

	Model1	Model2	Model3	Model4	Model5	Average
MSE	43.14	46.46	44.65	45.08	44.90	44.85
MAE	5.29	5.65	5.38	5.40	5.37	5.42
MAPE	0.85	0.93	0.84	0.84	0.86	0.86

MSE: mean squared error; MAE: mean absolute error; MAPE: mean absolute percent error.

By setting the hyperparameters of five groups of models and training the GBRT model respectively, the prediction results of the test set are obtained. After five groups of prediction results, including mean squared error (MSE) is 44.85, mean absolute error (MAE) is 5.42 and mean absolute percent error (MAPE) is 0.86.

3.3. Importance analysis of predictive variables

Figure 2 shows the ranking results of the importance of various features after the model converges. The experimental results showed that the clinical indicators contributed more to the model. Age and

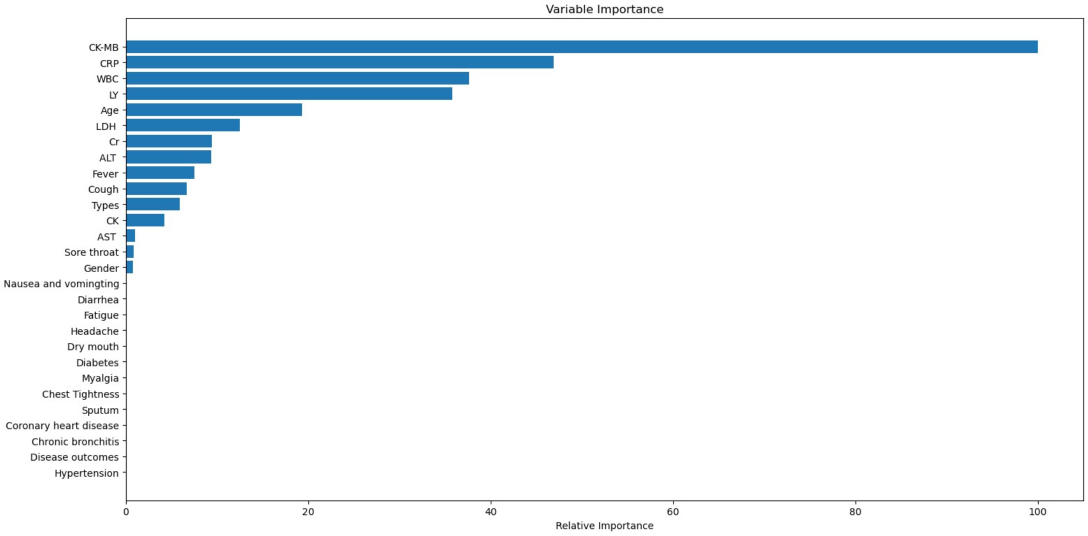

Figure 2. Ranking results of importance of various features to GBRT model.

gender, as intrinsic characteristics of the patients, also played a positive role in promoting the model predictions. However, in addition to patient type, feeder and sore throat, the pathological symptoms of patients in hospitals contribute little to the prediction of the model.

Through comprehensive analysis of Table 6 and Figure 2, we find that the model has a certain prediction ability for the LOS of patients, but the model still has some room for improvement. According to the ranking of feature importance, 13 features with low contribution are eliminated. In addition, we review Table 2 and find that the length of hospitalization of the data is too large. Due to strict policy control and the first encounter with such diseases in the hospital, patients may be required to stay for a few more days. Therefore, the discharge duration of each hospitalized patient, and the results are shown in Figure 3. The length of stay of hospitalized patients has a decreasing trend. The length of stay of patients admitted in July was significantly higher than that of patients admitted in August. According to the statistics, the average length of stay of 29 patients admitted in July was 26.6 days, and the median length of stay was 27 days; the average length of stay of 137 patients admitted in August was 16.07 days, and the median length of stay was 16 days. The length of stay of patients in August was closer to the average length of stay in China mentioned in the literature [12]. According to the statistical results, we consider excluding the data records in July and only using the patient data in August for modeling. Finally, our data set has 293 records and 15 features.

On the improved data set, five groups of super parameters of the GBRT model are found again, and the results are shown in Table 7.

The data set was 293 data records generated by 136 patients admitted in August, and 15 features were used for training. The data set is also divided by 8 : 2. The prediction results on the test set are shown in Table 8.

The experimental results in Table 8 show that the best MSE is 24.23, MAE is 4.16 and MAPE is 0.74. The performance indexes of the model have been improved to a certain extent.

Figure 4 shows the importance ranking results of various features of the improved model. We can see that clinical examination indexes such as CK-MB,CK-LDH and CRP have a very high contribution

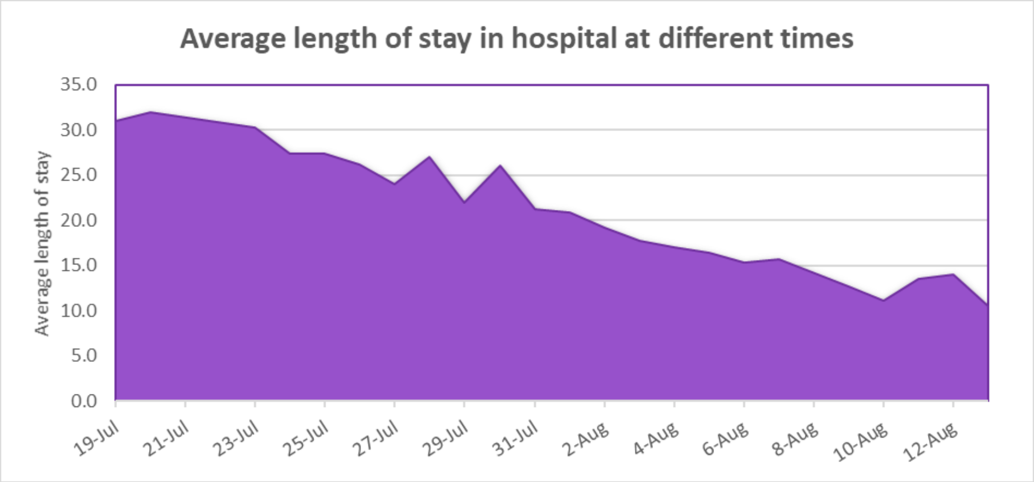

Figure 3. Average length of stay corresponding to different admission time points.

Table 7. Hyperparameters value of improved GBRT model.

Hyperparameters	Model6	Model7	Model8	Model9	Model10
learning rates	0.05	0.05	0.01	0.01	0.01
the depth of the tree	2	2	7	3	10
min samples split	5	5	5	5	2
min samples leaf	1	1	9	8	9
number of estimators	10	20	40	50	20

Table 8. Prediction results of improved GBRT model.

	Model6	Model7	Model8	Model9	Model10	Average
MSE	24.28	24.35	24.11	23.84	24.57	24.23
MAE	4.19	4.14	4.15	4.12	4.21	4.16
MAPE	0.73	0.76	0.73	0.76	0.73	0.74

to the prediction of the model. In addition, gender and age affect various states of the human body.

4. Discussion

The data used in this paper were produced from July to August 2020. On the one hand, the virus was more harmful to the body at that time; on the other hand, it was the first time the administrative department and medical unit in the place where the outbreak occurred to encounter the epidemic, so the hospitalization time of patients was longer than that mentioned in the literature [12]. Due to the small number of samples collected in this paper, this paper did not separate the validation set for hyperparameter selection, but directly divided the data into the training set and the test set by 8 : 2. Finally, the prediction results of five groups of models were shown, and the average value was used for reconciliation. In future work, we can try to further divide the training set, use the validation set to optimize the hyperparameters of the model, use the test set to estimate the prediction results of the model, and apply the cross-validation method to estimate the performance of the model more accurately

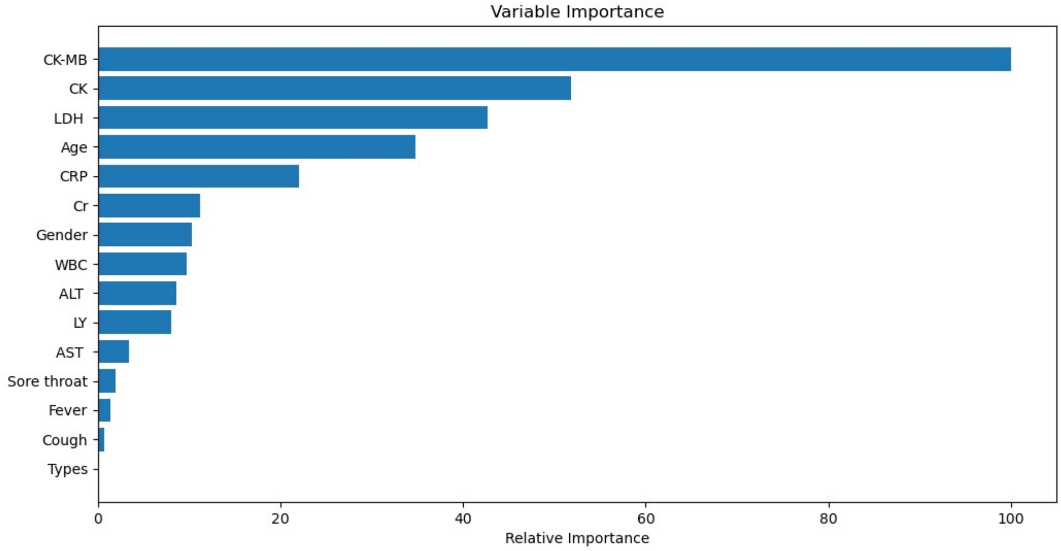

Figure 4. Ranking results of importance of various improved features to GBRT model.

when dividing the test set. In the aspect of data feature selection, machine learning algorithm is used in this paper, and all available features are taken as alternative features for training modeling. After the first round of iterative training, the features were sorted according to their importance, and the features with small contributions were eliminated. The interpretability of machine learning models has always been A concern for clinicians. Chia et al. [21] used the Cox-proportional hazards (CPH) model to screen features before modeling. CPH can provide clinicians with an alternative to interpret features. In future work, scholars can try to add a correlation analysis method at the very beginning to pre-select features. We are concerned that Pham et al. [22] mentioned that patients with a history of in-patient visits or if they received a high amount of treatment in their current visit were found more likely to be readmitted.

In future data collection work, secondary admission patients deserve follow-up attention. In addition, we note that the contribution of gender to model performance is relatively high. Li et al. [23] took 88,611 teachers as the research object, and the multiple logistic regression model was used to analyze the anxiety state during the epidemic. The results of the experiment showed that the anxiety of women was higher than that of men. We speculate that differences in anxiety due to sex may have some effect on the length of hospital stay.

The GBRT model is a typical algorithm in the integrated learning algorithm, which appears very frequently in the field of data analysis. In many machine learning competitions, parameter players often win the championship by relying on the integrated learning algorithm. In some tasks, their model performance may even be better than most deep learning algorithms. In future work, scholars can compare models through various methods, and they can also try neural network algorithm and complex deep learning algorithm. Although the neural network model consumes a lot of computing resources, the neural network has strong nonlinear mapping ability, strong robustness and strong self-learning ability. In a similar study, Ren et al. [24] in traffic flow forecasting, considering the interference of special events on short-term flow forecasting, a state-and-trend unit similarity degree (SD) measurement method and increment-based prediction model are proposed. Future research on LOS prediction can consider short-term trends and observation state information, and carry out dynamic fine-tuning strategies for the model to try to improve the performance of the model.

The model proposed in this paper also has some shortcomings. (1) In this paper, the data filling method was adopted to expand the data set, but the utilization rate of clinical index data generated by reexamination was still very low. (2) In the construction of the data set, replication padding was used for demographic characteristics and pathological features at admission. This method leads to a single data feature and is not conducive to model training (3) This paper only tries the traditional machine learning model. In the later stage, we can try the deep learning method to improve the prediction accuracy of the model.

5. Conclusions

In this paper, a prediction model for the length of hospital stay of COVID-19 patients based on GBRT was established. The constructed data set included demographic characteristics, clinical examination indicators, and pathological features at admission. The original data were 3141 records generated from 166 patients. Patients with excessive data record deletion, juvenile patients and patients admitted to the hospital before August were excluded, and 293 data records of 136 patients were finally retained. After super parameter selection and feature importance screening, the best results of MSE, MAE and MAPE were 23.84, 4.12 and 0.76, respectively. This model has a certain predictive ability and is helpful to medical management and decision aid. Finally, the importance ranking of data features is analyzed. Clinical indicators such as CK-MB, CK-LDH and CRP have a significant influence on the prediction of hospital stay.

Author contributions

Conceptualization: FL. Methodology: ZZ and TZ. Software: ZZ and XT. Validation: ZZ, TZ and GM. Formal analysis: ZZ and ZL. Investigation: ZZ, TZ, GM, YW and ZL. Data Curation: ZZ, TZ, GM, YW and ZL. Writing - Original Draft: ZZ. Writing-Review and Editing: TZ and YS. Supervision: FL. Project Administration: FL. Funding acquisition: FL. All authors critically read the manuscript and gave final approval for publication.

Acknowledgements

This research was supported by the Natural Science Foundation of Xinjiang (2022D01C183, 2021D01C268).

Conflict of interest

The authors declare no conflict of interest.

References

1. Y. Zhang, The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China, *China CDC Wkly*, **2** (2020), 113–122.

2. K. Li, C. Zhang, L. Qin, C. Zhang, A. Li, J. Sun, et al., A nomogram prediction of length of hospital stay in patients with COVID-19 pneumonia: A retrospective cohort study, *Dis. Markers*, **2021** (2021). https://doi.org/10.1155/2021/5598824

3. R. Yaesoubi, S. You, Q. Xi, N. A. Menzies, A. Tuite, Y. H. Grad, et al., Simple decision rules to predict local surges in COVID-19 hospitalizations during the winter and spring of 2022, preprint, arXiv: 2021.12.13.21267657. https://doi.org/10.1101/2021.12.13.21267657

4. X. Chen, W. Gao, J. Li, D. You, Z. Yu, M. Zhang, et al., A predictive paradigm for COVID-19 prognosis based on the longitudinal measure of biomarkers, *Briefings Bioinf.*, **22** (2021). https://doi.org/10.1093/bib/bbab206

5. L. Wynants, B. Van Calster, M. Bonten, G. Collins, T. Debray, M. De Vos, et al., Prediction models for diagnosis and prognosis of covid-19 infection: Systematic review and critical appraisal, *Br. Med. J.*, **369** (2020). https://doi.org/10.1101/2020.03.24.20041020

6. C. Bardelli, Inference on COVID-19 epidemiological parameters using bayesian survival analysis, *Entropy*, **23** (2021). https://doi.org/10.3390/e23101262

7. Y. You, M. Chen, X. Chen, W. Yu, Diaphragm thickness on computed tomography for nutritional assessment and hospital stay prediction in critical COVID-19, *Asia Pac. J. Clin. Nutr.*, **31** (2022), 33–40.

8. T. Chiam, K. Subedi, D. Chen, E. Best, F. B. Bianco, G. Dobler, et al., Hospital length of stay among COVID-19-positive patients, *J. Clin. Transl. Res.*, **7** (2021), 377–385.

9. M. G Usher, R. Tourani, G. Simon, C. Tignanelli, B. Jarabek, C. E Strauss, et al., Overcoming gaps: Regional collaborative to optimize capacity management and predict length of stay of patients admitted with COVID-19, *J. Am. Med. Inf. Assoc.*, **4** (2021), ooab055. https://doi.org/10.1093/jamiaopen/ooab055

10. E. Ahlstrand, S. Cajander, P. Cajander, E. Ingberg, E. Löf, M. Wegener, et al., Visual scoring of chest CT at hospital admission predicts hospitalization time and intensive care admission in Covid-19, *Infect. Dis.*, **53** (2021), 622–632. https://doi.org/10.1080/23744235.2021.1910727

11. A. Lasbleiz, B. Cariou, P. Darmon, A. Soghomonian, P. Ancel, S. Boullu, et al., Phenotypic Characteristics and Development of a Hospitalization Prediction Risk Score for Outpatients with Diabetes and COVID-19: The DIABCOVID Study, *J. Clin. Med.*, **9** (2020), 3726. https://doi.org/10.3390/jcm9113726

12. C. Eastin, T. Eastin, Clinical characteristics of coronavirus disease 2019 in China, *J. Emerg. Med.*, **58** (2020), 711–712. https://doi.org/10.1016/j.jemermed.2020.04.004

13. Q. J. Leclerc, N. M. Fuller, R. H. Keogh, K. Diaz-Ordaz, R. Sekula, M. G. Semple, et al., Importance of patient bed pathways and length of stay differences in predicting COVID-19 hospital bed occupancy in England, *BMC Health Serv. Res.*, **21** (2021). https://doi.org/10.1101/2021.01.14.21249791

14. J. Ebinger, M. Wells, D. Ouyang, T. Davis, N. Kaufman, S. Cheng, et al., A Machine learning algorithm predicts duration of hospitalization in COVID-19 patients, *Intell. Based Med.*, **5** (2021),100035. https://doi.org/10.1016/j.ibmed.2021.100035

15. B. Mahboub, M. T. A. Bataineh, H. Alshraideh, R. Hamoudi, L. Salameh, A. Shamayleh, et al., Prediction of COVID-19 Hospital length of stay and risk of death using artificial intelligence-based modeling, *Front. Med.*, **8** (2021). https://doi.org/10.3389/fmed.2021.592336

16. A. Lopez-Cheda, M. A. Jacome, R. Cao, P. M. De Salazar, Estimating lengths-of-stay of hospitalised COVID-19 patients using a non-parametric model: A case study in Galicia (Spain), *Epidemiol. Infect.*, **149** (2021), e102. https://doi.org/10.1017/S0950268821000959

17. A. Henzi, G. Kleger, M. P. Hilty, P. D. W. Garcia, J. F. Ziegel, Probabilistic analysis of COVID-19 patients' individual length of stay in Swiss intensive care units, *Plos One*, **16** (2021). https://doi.org/10.1371/journal.pone.0247265

18. T. Dan, Y. Li, Z. Zhu, X. Chen, W. Quan, Y. Hu, et al., Machine Learning to Predict ICU Admission, ICU mortality and survivors' length of stay among COVID-19 patients: Toward optimal allocation of ICU resources, in *2020 IEEE International Conference on Bioinformatics and Biomedicine*, (2020), 555–561. https://doi.org/10.1109/BIBM49941.2020.9313292

19. H. Yue, Q. Yu, C. Liu, Y. Huang, Z. Jiang, C. Shao, et al., Machine learning-based CT radiomics model for predicting hospital stay in patients with pneumonia associated with SARS-CoV-2 infection: A multicenter study, *Ann. Transl. Med.*, **8** (2020), 859. https://doi.org/10.21037/atm-20-3026

20. D. Rozenbaum, J. Shreve, N. Radakovich, A. Duggal, L. Jehi, A. Nazha, Personalized prediction of hospital mortality in COVID-19-positive patients, *Mayo Clin. Proc. Innovations Qual. Outcomes*, **5** (2021), 795–801. https://doi.org/10.1016/j.mayocpiqo.2021.05.001

21. A. H. T. Chia, M. S. Khoo, A. Z. Lim, K. E. Ong, Y. Sun, B. P. Nguyen, et al., Explainable machine learning prediction of ICU mortality, *Elsevier Ltd*, **25** (2021), 100674. https://doi.org/10.1016/j.imu.2021.100674

22. H. N. Pham, A. Chatterjee, B. Narasimhan, C. W. Lee, D. K. Jha, E. Y. F. Wong, et al., Predicting hospital readmission patterns of diabetic patients using ensemble model and cluster analysis, in *2019 International Conference on System Science and Engineering (ICSSE)*, (2019), 273–278. https://doi.org/10.1109/ICSSE.2019.8823441

23. Q. Li, Y. Miao, X. Zeng, C. S. Tarimo, C. Wu, J. Wu, Prevalence and factors for anxiety during the coronavirus disease 2019 (COVID-19) epidemic among the teachers in China, *J. Affect. Disord.*, **277** (2020), 153–158. https://doi.org/10.1016/j.jad.2020.08.017

24. Y. Ren, H. Jiang, N. Ji, H. Yu, TBSM: A traffic burst-sensitive model for short-term prediction under special events, *Knowl-Based Syst.*, **240** (2020), 108120. https://doi.org/10.1016/j.knosys.2022.108120

AIMS Press